CASE STUDY FOR CREDIT ASSIGNMENT USING EDA

Submitted By:
Ankita Chaudhary

PROBLEM STATEMENT & BUSINESS OBJECTIVES

This case study aims to identify patterns which indicate:

- If a client has difficulty paying their instalments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

PROBLEM STATEMENT

1. PROBLEM STATEMENT UNDERSTANDING
2. DATA UNDERSTANDING (APPLICATION DATA)
3. DATA CLEANING: IDENTIFICATION OF MISSING VALUES AND TREATMENT
   OUTLIER AND TREATMENT
4. DATA IMBALANCE
5. SANITY CHECKS
6. UNIVARIATE ANALYSIS --> ANALYZING ONE VARIABLE AT A TIME : HISTOPLOT
7. BIVARIATE ANALYSIS --> ANALYZING TWO VARIABLES AT A TIME : BARPLOT
8. ANALYSIS OF PREVIOUS APPLICATION DATA SET WITH ABOVE STEPS

DATA UNDERSTANDING (APPLICATION DATA)

Checking the Shape, Info, D-Types, Size and Describe of the Application Data and Previous Data to get a quick understanding of the data.

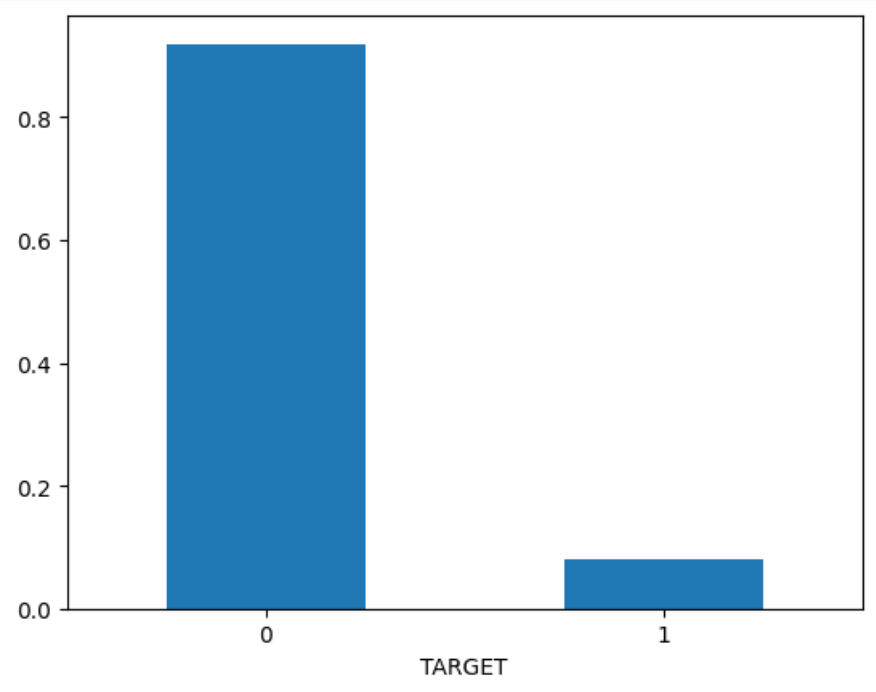
DATA CLEANING

- Identification of Missing Values and Treatment:
- There were several Columns with missing value percentage greater than 40% so we dropped them.
- Remaining columns with missing values we imputed them with mean/median/mode as required.
- Outliers: There were several columns which have outliers present in them.
- Depicted through boxplot in columns named as :
- AMT_ANNUITY
- AMT GOODS_PRICE
- NAME_TYPE_SUITE
- OCCUPATION_TYPE

DATA IMBALANCE

When we observed the data set it was highly imbalanced almost 91.9% clients were Non Payment Difficulties (0) and about 8.1% for Payment Difficulties(1).

0= Non-Payment Difficulties
1= Payment Difficulties



SANITY CHECKS

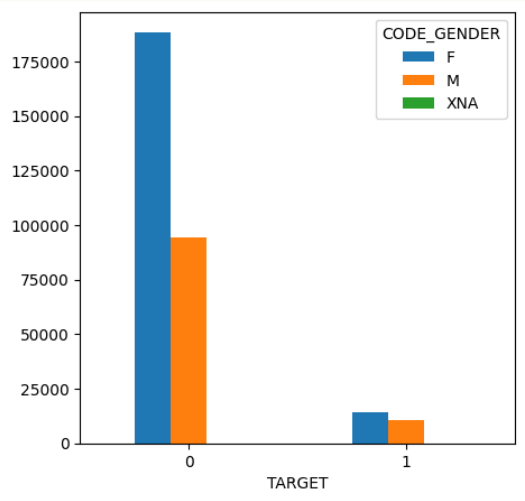
- Doing analysis through the data observed some columns have negative values. So to correct them we did sanity checks in these columns named as:

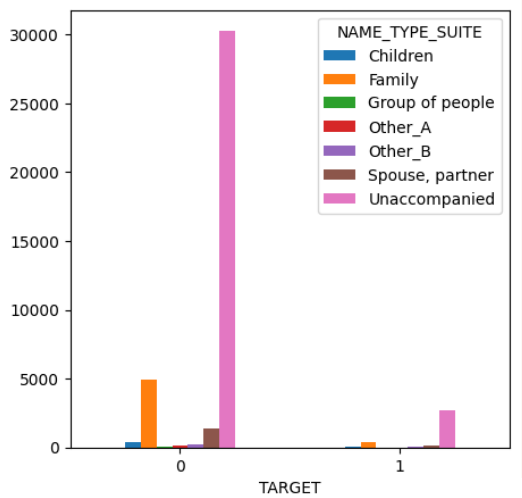
- DAYS BIRTH", "DAYS_EMPLOYED", "DAYS_REGISTRATION", "DAYS_ID_PUBLISH","DAYS_LAST_PHONE_CHANGE

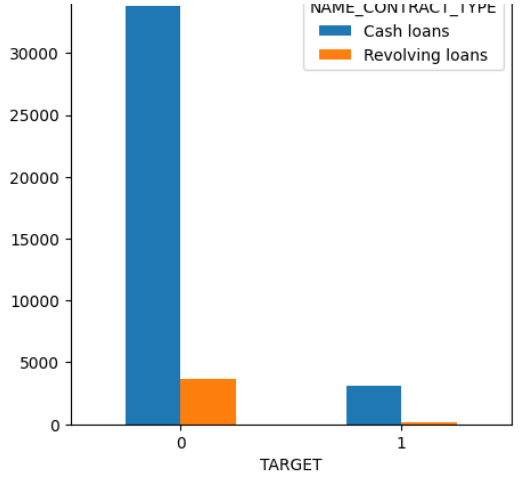
BIVARIATE ANALYSIS

Conclusion from analysis:

- We can say that most client are females, working, have secondary education and are married.
- We can say taking loan from the bank someone accompanied with them as well.





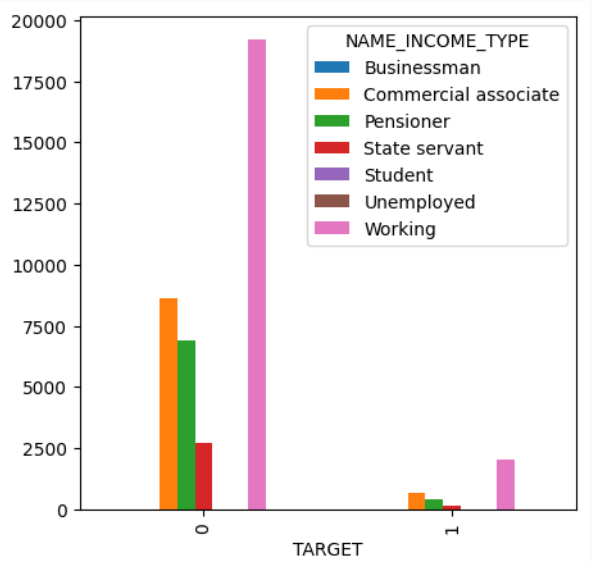


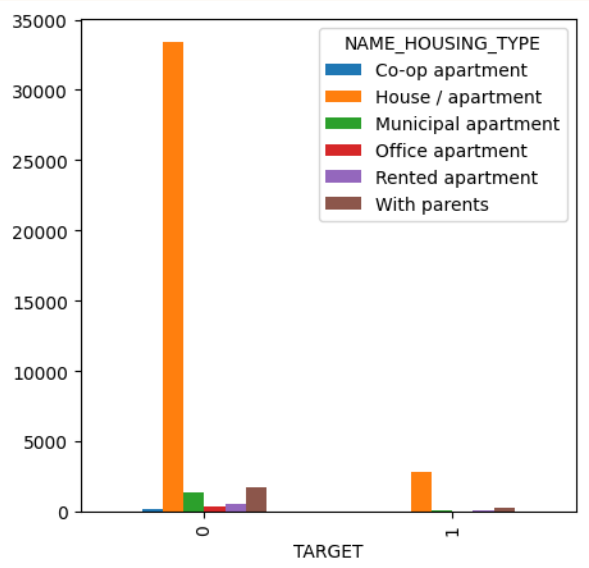
BIVARIATE ANALYSIS

CONCLUSION FROM ANALYSIS:

- WE OBSERVED THAT THERE IS SHARP INCREASE IN LABOURERS CATEGORY IN PAYMENT

- DEFAULT CATEGORY AND ALSO MALE LABORER HAVE MOST DIFFICULTIES IN PAYING LOAN.





INSIGHTS

- There exists more clients who changed their registration details after 4000 days of approval of loan.
- For few not default clients, time taken to publish id's are higher than default clients.
- The application process start hours taken for default and not default cases are similar.
- In non default cases, people keep their phone numbers for greater time.
- People with greater number of days born count are less likely to default.
- In non default case AMT_GOODS PRICE contains more outliers than default case.
- In default case, most of the clients amount annuity tends to be greater than 25000(median value).
- Whose credit amount is greater than 50000 tends to be less default than compared to default cases and vice versa.
- people with higher no of employment days are less likely to default.
- Majority of defaulting people are having less total income.

INSIGHTS

As we can see from graphs:

- High number of applications are filed in 9 AM to 2 PM for both Current and Previous data.
- So busiest hours for bank are form 9 AM to 2 PM.
- nuclear family tends to take more loans.
- Previously bank had high unused offers but currently refused is high incase of AMT_GOODS_PRICE.
- Previously bank had high unused offers and currently cancelled/refused offers are similar for AMT_ANNUITY.
- Previously bank had high unused offers and currently high number of refused offers for AMT_CREDIT.

INSIGHTS

- AMT_CREDIT_Previous has highest refused cases and AMT_CREDIT_Current is similar for all 4 cases.
- time spent in unused offer is higher as compared to other categories.
- So bank should reduce time spent on unused offer.
- Nuclear family(2-3 people in family) get highest approval.
- Previously most of the applications were cancelled or refused
- But now Refused/Cancelled/Approved/Unused all four have similar situation for AMT_GOODS_PRICE.
- Previously most of the applications were cancelled or refused
- But now Refused/Cancelled/Approved/Unused all four have similar situation for AMT_ANNUITY.

RECOMMENDATIONS

Target/focused variable for Application dataset - **TARGET**

Target/focused variable for Previous dataset - **NAME_CONTRACT_STATUS**

Top Major variables to consider for loan prediction:

1. NAME_EDUCATION_TYPE
2. AMT_INCOME_TOTAL
3. DAYS_BIRTH
4. AMT_CREDIT
5. DAYS_EMPLOYED
6. AMT_ANNUITY
7. NAME_INCOME_TYPE
8. CODE_GENDER
9. NAME_HOUSING_TYPE

The above mentioned variables are to be considered before approving application to minimize risk of loss.

THANK YOU